Physical Activity Recognition using Posterior-adapted Class-based Fusion of Multi-Accelerometers data

Alok Kumar Chowdhury, Dian Tjondronegoro, *Sr. Member, IEEE*, Vinod Chandran, *Sr. Member, IEEE,* and Stewart G. Trost

***Abstract* – This paper proposes the use of posterior-adapted class-based weighted decision fusion to effectively combine multiple accelerometers data for improving physical activity recognition. The cutting-edge performance of this method is benchmarked against model-based weighted fusion and class-based weighted fusion without posterior adaptation, based on two publicly available datasets, namely PAMAP2 and MHEALTH. Experimental results show that: (a) posterior-adapted class-based weighted fusion outperformed model-based and class-based weighted fusion; (b) decision fusion with two accelerometers showed statistically significant improvement in average performance compared to the use of a single accelerometer; (c) generally, decision fusion from 3 accelerometers did not show further improvement from the best combination of 2 accelerometers, (d) a combination of ankle and wrist located accelerometers showed the best overall performance compared to any combination of two or three accelerometers.**

***Index Terms*—Activity Recognition, Accelerometer, Decision Fusion, Class-Based Weighted Fusion**

I. Introduction

Physical inactivity is a critical health risk factor [1], which triggers the need for real time physical activity (PA) recognition and quantification of the frequency and intensity of each PA instances using accelerometer-based motion sensors [2, 3]. A range of approaches including rule-based (such as threshold/hierarchical), supervised and unsupervised classification algorithms have been proposed for PA recognition [4-8]. The choice between using machine learning or rule-based approach is often determined by the availability of a suitable training set. In the case of data scarcity, rule based systems are usually used based on the domain knowledge. Most recent papers in the PA domain suggested the use of supervised machine learning algorithms [9, 10] as there are usually enough labelled data to train a reliable machine-learning model. However, previous studies generally used their own datasets, with no validation of results across variations of datasets, size of datasets and activity type selections. Therefore, the performances of existing algorithms have been found to be inconsistent and dependent on the sample used to generate the training data and the activity targets under investigation [5].

Multiple accelerometers placed at different body locations has been found to be effective in improving the accuracy of PA recognition and the performance depends on activity type [3, 11, 12]. Acceleration data from multiple locations can be combined using feature- or a decision-level fusion approach [13]. Decision-level fusion has been found to be more accurate than feature fusion in other domains [14]; however, it has not been systematically investigated for PA recognition.

This paper's key contribution is to propose the use of posterior-adapted class-based weighted decision fusion. It is novel, as class-based decision fusion has not been used for PA recognition, while it has been found to perform better than model-based decision fusion [15]. Moreover, using posterior probability of the test data can further improve the performance and it has also not been utilized in PA domain. In model-based fusion, a model is developed for each accelerometer location, then the fusion assigns a weight for each model based on the overall performance based on its training data. Such approach is theoretically less robust compared to class-based fusion, which focuses on the class (i.e. activity) wise performance of the models. Posterior-adaptation means that the class-based weights are dynamically adjusted using the confidence scores from each classification model based on real observations (i.e. test data).

Aside from finding the most effective fusion technique, another challenge is to determine the best combination of sensor placements for optimal PA recognition. Therefore, our experiments have investigated how decision-level fusion can optimally combine multiple classification models, where each model is trained using the accelerometer data obtained from ankle, chest and wrist respectively. The robustness of our proposed method has been tested against two publicly available datasets and benchmarked with model-based and class-based weighted decision fusion techniques. To sum up, the novelty of this paper is proposing the use of posterior-adapted class-based weighted decision fusion to effectively

Manuscript received _____; revised _____; accepted _____. Date of publication _____; date of current version _____.

Alok Kumar Chowdhury, Dian Tjondronegoro, and Vinod Chandran are with Queensland University of Technology, Science and Engineering Faculty, 2 George St, Brisbane, Australia. (email: alok.chowdhury@qut.edu.au, dian@qut.edu.au, vinod.chandran@ieee.org)

Stewart G. Trost is with Queensland University of Technology, Institute of Health and Biomedical Innovation at QLD Centre for Children's Health Research, School of Exercise and Nutrition Sciences, 2 George St, Brisbane, Australia (e-mail: s.trost@qut.edu.au).

combine multiple accelerometers data for improving physical activity recognition.

II. Related Work

PA recognition accuracy has been found to be dependent on the accelerometer locations and types of PAs. For example, Atallah, et al. [16] used k nearest neighbor (KNN) classifier and Bayesian classifier and found that the wrist location was good for recognizing very low-intensity-level and medium-intensity-level activities. For low-intensity-level and transition activities, the waist location was the best. However, the authors did not combine data from the accelerometers to find the optimal combinations.

Some studies compared the performance of classifiers trained on data from the combination of different accelerometer locations. Bao et al. [17] used feature fusion on the accelerometer data collected from the upper-arm, lower-arm, hip, thigh, and ankle and then applied several learning algorithms. They found a decision tree to be the best performer (84%) when all sensors were fused, while the combination of thigh and wrist accelerometer provided 3% less accuracy. However, the authors did not investigate all possible accelerometer location combinations or the effect of accelerometer location on recognizing different PA types. A comprehensive study by Cleland et al [18] used feature fusion and compared the performance of support vector machine classifiers trained on accelerometer data from six body locations (lower back, wrist, foot, chest, hip, and thigh) and their combinations. Compared to a single accelerometer, combining data from any two locations resulted in a significant improvement in performance. However, combining data from three or more accelerometers provided no further improvements in performance. Kern, et al. [19], Gjoreski et al [20] and Olguin et al. [21] also reported significant improvements in recognition performance when combining two or more accelerometer locations. Notably, all of the aforementioned studies used feature fusion, which is more prone to noisy and redundant data compared to decision fusion approach [13]. These studies did not use multiple public datasets and consider the varying performance of single accelerometers for different activities when combining different accelerometer positions.

There are existing studies in pattern recognition that investigated the best classifier combination for decision fusion, such as using a diversity measure analysis [22]. In activity recognition, the commonly used decision fusion rules include majority voting, summation, hierarchical fusion and Bayesian fusion [23]. Banos, et al. [24] proposed hierarchical-weighted decision fusion by combining the advantages of the hierarchical decision and majority voting models which utilized class-level classifiers and sensor-level classifier for making decisions. While several weighted fusion techniques (classifier, class and sample-based) were compared empirically in [15], class-based fusion seems to be more suitable for accelerometer fusion in PA recognition due to the variation in the class-wise performance of different placement of accelerometers. However, class-based fusion is yet to be fully investigated in the PA domain. Moreover, the approach for calculating the weights in class-based fusion needs improvement as it uses training errors to evaluate testing reliability. Adaptation of class-based weights using the posterior probability of the test instances should further improve the fusion techniques. Zhang and Zhang [25] showed that adjusting the probabilities derived from the training output confusion matrix using the decision reliability can improve the decision-making accuracy. However, they did not use class-based weights, and did not apply their fusion algorithm for the PA recognition.

III. Methods

The framework comprises pre-processing, feature extraction, normalization, feature selection, and classification. These steps were simultaneously applied to data from each accelerometer location (e.g. ankle, chest, and wrist), resulting in activity candidates. The final decision (i.e. which activity is the most likely) was achieved by applying a posterior-adapted class-based weighted decision fusion. Each step will be described in this section.

A. Pre-processing

Each of the 3-axis (x,y,z) accelerometer data was converted to a time-series data structure. A linear interpolation method was used to impute missing data in the middle of a labelled activity sequence. The missing values at the end of each labeled activity sequence were replaced by the previous value.

B. Feature Extraction

For each of the 3-axis accelerometer data, a set of 45 features (in time- and frequency- domain) was extracted from a 2-second sliding window without overlapping. Short windows (interval 1–2 second) were used, as it has been shown to demonstrate the best trade-off between accuracy and speed in PA recognition [26]. Specifically, 2-second window was empirically set as it was capable of capturing the periodic movements for the selected PA classes.

Table 1 lists the extracted features from each window. These features were combined from the features extracted in previous PA recognition studies [17, 27-30].

TABLE 1
LIST OF FEATURES EXTRACTED FROM EACH WINDOW OF AN ACCELEROMETER.

No	Features	Feature Count
1	Mean for each axis of a 3-axis accelerometer	3
2	Standard deviation for each axis of accelerometer	3
3	Minimum value for each axis	3
4	Maximum value for each axis	3
5	Variance for each axis	3
6	Median value for each axis	3
7	Skewness for each axis	3
8	Kurtosis for each axis	3
9	Energy for each axis	3
10	Cross-correlation of accelerometer axis	3
11	Principal frequency for each axis	3
12	Magnitude of principal frequency for each axis	3
13	Median crossing for each axis	3
14	25^{th} percentile for each axis	3
15	75^{th} percentile for each axis	3
	Total number of features extracted	**45**

C. Normalization & Feature Selection

Normalization is required to limit feature values within a range, and in this case, we set the range to zero mean and unit variance using linear methods. For example, a feature x can be normalized using following formula (1).

$$\hat{x}_i = \frac{x_i - \overline{x}}{\sigma} \quad (1)$$

Where, $\overline{x}$ and σ are mean and standard deviation respectively.

Use of unnecessary features may lead to over-fitting, low performance and computational load [31]. Therefore, instead of adopting all the 45 features for classification, correlation-based feature selection method was adopted to select the most useful features. This feature selection method is fast, simple, and found to be useful in previous studies [32]. In this study, the training data was used to compute the correlations between each labelled activity and feature. Features that have a correlation of 0.25 or greater (threshold was suggested in [33]) were selected for training and testing the classifiers.

D. Classification Algorithms

In order to find the best classification approach, several state-of-the-art machine learning methods were initially benchmarked, including binary decision tree (BDT), support vector machine (SVM), and deep neural network (DNN), random forest (RF) and Adaboost.

In our implementation, the maximum decision split for BDT was set to 20. The DNN used two auto-encoders to convert inputs into 35 and 20 deep features respectively. A softmax layer was trained using 20 deep features for activity classification. To reduce overfitting, L2-weight regularization (value set to 0.001) was added to train criterion. In RF, random subset of predictors for each decision split was equal to the square root of the total number of available features. For the Adaboost.M2 algorithm, a multi-class classification method with 100 learning cycle was implemented.

Based on the experimental results (see section V for details), SVM was selected as the best classification algorithm, as it showed the highest classification accuracy compared to other classification algorithms, although the difference was marginal.

E. Decision Fusion Techniques

Let's consider, the fusion of decisions from n models for a m-class problem. The sets of models and classes can be presented as $M = \{M_1, M_2 \ldots M_n\}$ and $C = \{C_1, C_2 \ldots C_m\}$. When classifying a test instance (x), each model provides a predicted class label along with a posterior probability of the predicted label, which is a measure of the confidence of the decision from that model for that test instance. Let the predicted vector for that instance be $V(x) = \{V_1(x), V_2(x) \ldots V_n(x)\}$ where each $V_i(x) \in C$, and the posterior probabilities be $W_2(x) = \{W_{21}(x), W_{22}(x) \ldots W_{2n}(x)\}$. A decision fusion technique provides a final prediction for x by combining individual predictions $\{V(x)\}$.

A *model-based weighted voting* assigns a weight to each model/classifier based on the overall performance of that model on the training data irrespective of classes [13, 15]. This weight is independent of its predicted class. In the fusion step, a weighted majority is used to decide the final predicted class. In contrast, a *class-based weighted decision fusion* assigns weights to all classes based on the prior knowledge of the model's prediction performance for the different classes [15]. In the fusion step, a weighted majority is again used to decide the final predicted class but the weights are now different and the majority class may be different. This study proposes *posterior-adapted class-based fusion*, which adjusts the class-based weights for each test instance using the posterior probability of the model on the prediction. The steps to achieve these weighted decision fusion schemes are described below.

Weight calculation – Using 10-fold cross validation on the training data, both predicted training classes and true training classes are compared and the F1-scores for all classes are computed. F1-scores indicate the model's confidence for each class based on the training data, which are used as class-based weights. The 10-fold validation allows reliable calculation of expected class-wise performance on unseen data and avoid overfitting.

Let the class-based weights for models be $W_1 = \{W_{11}, W_{12} \ldots W_{1n}\}$, where W_{1i} is a collection of weights for all (m) classes for the i^{th} model, i.e. $\{w_{1i1}, w_{1i2} \ldots w_{1im}\}$. For model-based fusion, a weight for each model is calculated by taking average of class-based weights W_{1i}.

$$W_{avg\,i} = \overline{W_{1i}} \qquad 1 \le i \le n \quad (2)$$

Weight adjustment – For each test instance, the class-based weights are adjusted using the posterior probability of the predicted label. Let the adjusted class-based weights for the given test instance are $W_i(x) = \{w_{i1}, w_{i2} \ldots w_{im}\}$ $1 \le i \le n$. At first, the adjusted class-based weights are initialized to the class-based weights.

$$W_i(x) = W_{1i} \qquad 1 \le i \le n \quad (3)$$

Then, weights are adjusted by the posterior probability using the following equation.

$$w_{ik} = (\alpha * w_{ik} \underset{V_i = C_k}{+} (1-\alpha) * W_{2i}) \quad 1 \le k \le m,\ 1 \le i \le n \quad (4)$$

Here, α is a weight adjustment parameter, within 0 to 1, that requires tuning.

Model-based Fusion – This fusion scheme takes a weight ($W_{\text{avg i}}$) for each model and current prediction vector $\{V(x)\}$ to make a final prediction for a test instance (x). It computes the score for each of the predicted label by summing up the corresponding model's weight using equation (8).

$$Score_{V_i(x)} = \sum w_{avg\,i} \qquad 1 \le i \le n \quad (5)$$

Then it selects that predicted label $\{V_i(x)\}$ as final decision, which has the highest score.

Class-based Fusion – This fusion scheme considers the class-based weights $W_I(x)$ and current prediction vector $\{V(x)\}$ to make a final prediction for a test instance (x). It calculates score for each class using following formula.

$$Score_k = \sum_{V_i(x)=C_k} w_{1ik} \qquad 1 \le k \le m,\ 1 \le i \le n \qquad (6)$$

Finally, it selects the class label as final prediction, which has maximum score using equation (7).

$$final_{label} = C_{arg\,max_{k=1}^{m}\, Score_k} \qquad (7)$$

Posterior-adapted Class-based Fusion – This fusion scheme is similar to class-based fusion, but it used adjusted class-based weights $W_i(x)$. It calculates score for each class using the following formula.

$$Score_k = \sum_{V_i(x)=C_k} w_{ik} \qquad 1 \le k \le m,\ 1 \le i \le n \qquad (8)$$

Finally, it selects the class label as final prediction, which has maximum score using equation (7).

IV. Experiment

A. Datasets

Two publicly available PA monitoring datasets were chosen for the study, as they both have accelerometer sensors data from three body positions (ankle, chest and wrist). These data had been shown by previous studies [2, 34-36] to be effective for machine learning purposes, which confirms that there was enough data to train the machine learning models.

The **PAMAP2 Dataset** includes data from nine participants (1 female, 8 male), with age and body mass index (BMI) of 27.2 ± 3.3 years and 25.1 ± 2.6 kg/m2 respectively. Participants wore three Colibri wireless IMUs on their dominant-side wrist, ankle, and chest, when performing physical activities including lying down, sitting, standing, walking, running, cycling, Nordic walking, ascending stairs, descending stairs, vacuum cleaning, ironing clothes and jumping rope. Each sensor contains two three-dimensional (3D) acceleration sensor (scale: ±6g and ±16g) with a resolution of 13 bits, a gyroscope sensor, a magnetometer sensor, temperature, orientation and heart rate monitor sensors. The sampling rate of recorded acceleration data is 100 Hz. Further details of the study protocol can be found in [2, 36].

The **MHEALTH Dataset** includes data from ten participants, in an out-of-lab environment, while performing twelve physical activities. The physical activities include: standing still (1 min), sitting and relaxing (1 min), lying down (1 min), walking (1 min), climbing stairs (1 min), waist bends forward (20x), frontal elevation of arms (20x), knees bending (crouching) (20x), cycling (1 min), jogging (1 min), running (1 min), and jumping front & back (20x). During the data collection, Shimmer2 (Shimmer 2R, Real-time Technologies, Dublin, Ireland) wearable sensors were attached to the subject's chest, right wrist and left ankle. These sensors monitor 3D acceleration data (±6g) from chest, ankle, & wrist, electrocardiography (ECG) signal, 3D gyroscope data from ankle, & wrist, and 3D magnetometer data from ankle, & wrist. The sampling rate of recorded data is 50Hz. Further details on the data collection can be found in [34, 35].

Both datasets are fully labelled with each raw acceleration signal annotated based on the performed activity. For the purpose of this study, a subset data was extracted from both datasets. For PAMAP2, the selected activity classes were lying down, sitting, standing, walking, running, cycling, ascending stairs, and descending stairs. For MHEALTH dataset, lying down, sitting and relaxing, standing still, walking, running, cycling, climbing stairs, and jogging activities were chosen for analysis.

B. Implementation of the Framework

Figure 1 shows how the framework had been implemented. For each accelerometer location data (ankle, chest, and wrist), SVM classifier was applied in the four-phase processes: (1) **training phase**, where the classifier was trained using training data; (2) **weight calculation phase**, where the classifier was evaluated using 10-fold cross-validation of the training data and the resultant class-based weights and average weights were assigned, (3) **individual model decision phase**, where the classification model (trained in phase 1) was applied to a new/testing data and predicted label and its posterior probability, (4) **class-based weight adjustment phase**, where class-based weights from training data (output of phase 2) were adjusted using the posterior probability of the predicted label (output of phase 3), called *adjusted class-based weights.*

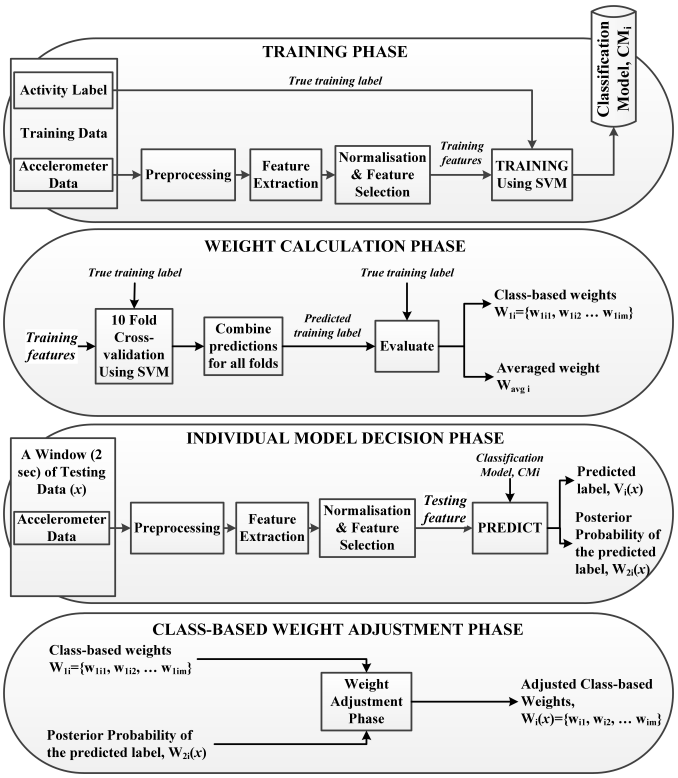

Fig. 1. Overview of the system developed for implementing the framework

Finally, a decision fusion phase (figure 2) combined the decisions from each individual sensor location using posterior-

adapted class-based weighted decision fusion, and also using model-based, class-based decision fusion techniques for benchmarking purposes.

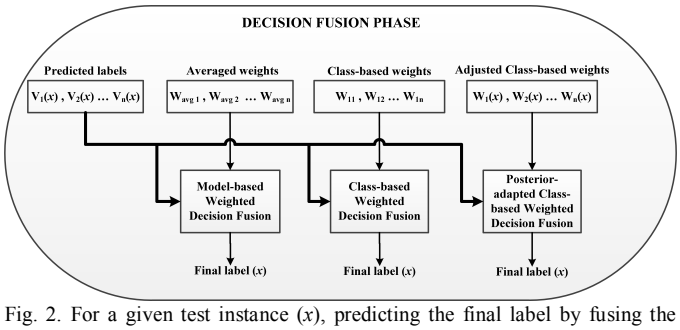

Fig. 2. For a given test instance (x), predicting the final label by fusing the decisions from accelerometer sensors using weights

C. Evaluation Approach and Metrics

Leave-one-subject-out cross-validation was used to evaluate and compare the classification models. This evaluation uses one subject's data for testing and remaining subject's data for training to conduct a subject-independent evaluation. Thus, all subject's data are considered once for testing (as suggested in [37]). In a real-world context, it is desirable for an activity recognition system to perform well for a new subject.

The performance of each classifier was evaluated by calculating precision, recall and F1-score. For each class, predictions were compared to ground truth labels and the number of true-positives (TP), true-negatives (TN), false-positives (FP), and false-negatives (FN) were calculated. Precision measures the exactness of a classifier while recall can measure the completeness of classifiers. These can be calculated for a particular class using the following equations.

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

The F1-score is a balanced combination of both precision and recall can be measured using the following formula.

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall}\ x\ 100\ \% \quad (11)$$

The predicted classes for each subject were combined and a confusion matrix was derived from the complete set. Then using the confusion matrix, F1-scores for all activity classes were computed to get an insight into the model's performance for each class. Let the number of subjects and classes are n and m respectively. The classes can be presented as $C = \{C_1, C_2 \ldots C_m\}$. Given that, true and predicted classes are $\{T_1, T_2 \ldots T_n\}$ and $\{P_1, P_2 \ldots P_n\}$ respectively. Where, T_i and P_i are true and predicted classes for i^{th} subject and $T_i \in C$, $P_i \in C$. The F1-Scores were calculated using the following steps.

Step 1: $P = \cup_{i=1}^{n} P_i\,; \quad T = \cup_{i=1}^{n} T_i$

Step 2: $F1Score_{CLASS\_WISE}(T, P) = \{FS_1, FS_2, \ldots FS_m\}$

Here, FS_k is the overall F1-Score of k^{th} class across subjects.

V. Results and Discussion

A. Evaluation of Classification Algorithms

Table 2 shows the F1-scores of machine learning algorithms using both PAMAP2 and MHEALTH datasets. The results were not conclusive in terms of deciding the best classification approach. Both RF and SVM consistently showed better performance for all three accelerometer locations across both datasets. However, for the remaining analyses, this study adopted SVM, as it gave the highest F1-score (82.32%) when averaged over all placement locations and both datasets, which is consistent with previous work [18].

TABLE 2
AVERAGE F1-SCORES FOR EACH CLASSIFICATION MODEL ACROSS BOTH DATASETS

Classifier	PAMAP2			MHEALTH			Average Across Both Datasets
	Ankle	Chest	Wrist	Ankle	Chest	Wrist	
SVM	84.72	81.00	80.86	83.64	79.62	84.10	**82.32**
RF	84.88	77.57	77.91	83.55	83.16	86.35	**82.24**
BDT	77.56	71.95	74.42	79.36	80.49	87.22	**78.50**
DNN	78.17	79.38	76.12	87.30	77.87	86.88	**80.95**
Adaboost	79.68	79.02	76.79	86.08	82.99	86.68	**81.87**

B. Evaluation of Different Fusion Techniques

Figures 3 and 4 show the average classification performances of model-based, class-based and posterior-adapted class-based decision fusion across different accelerometer location combinations for the PAMAP2 and MHEALTH datasets respectively. Error bars in both figures present 95% confidence interval (CI). The weight adjustment parameter (α) in posterior-adapted class-based weighted fusion was set to 0, 0.25, 0.5, 0.75 and 1. An α = 0.5 adjusts the weights by taking the average of the class-based weights and posterior probabilities and provided the best performance for the optimal accelerometer combination (Ankle + Wrist). Hence, the results reported in this paper used α = 0.5.

In both datasets, the posterior-adapted class-based weighted fusion consistently provided the best average F1-Scores for all accelerometer combinations compared to that obtained using either model-based or class-based weighted fusion. While the performance of model-based fusion was poor for most two accelerometer combinations, the class-based fusion performed well in most situations (F1-Scores were higher than model-based but lower than posterior-adapted class-based).

With PAMAP2, the posterior-adapted class-based weighted fusion provided statistically significant improvement in performance compared to model-based fusion for all two accelerometer combinations, but not A+C+W. With MHEALTH, the posterior-adapted class-based weighted fusion provided statistically significant improvements in performance compared to model-based fusion for A+W and C+W, but not A+C or A+C+W. Across all accelerometer configurations, posterior-adapted class-based weighted fusion consistently provided higher classification accuracy than class weighted decision fusion; however, there were no statistically significant differences in average F1-Scores.

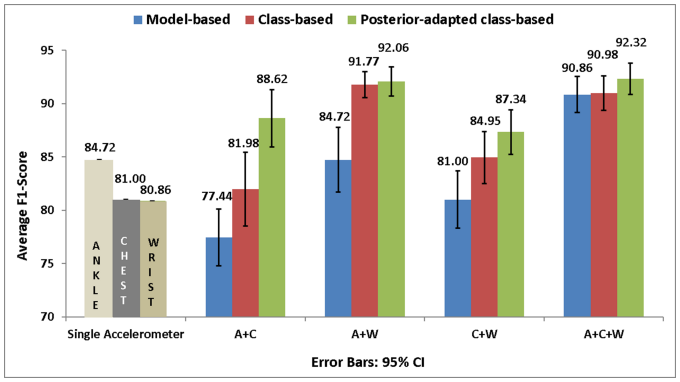

Fig. 3. Average F1-Score comparison for model-based, class-based and posterior-adapted class-based decision fusion with the PAMAP2 dataset

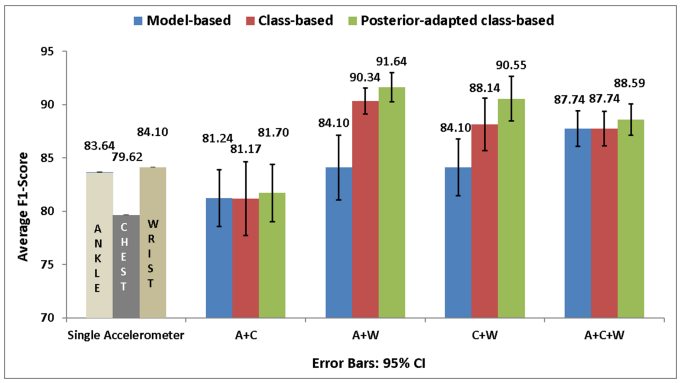

Fig. 4. Average F1-Score comparison for model-based, class-based and posterior-adapted class-based decision fusion with the MHEALTH dataset

C. *Activity-Wise Classification Performance*

Tables 3 and 4 report the class/activity-wise and average F1-scores for single location classifiers and all possible combinations of accelerometer locations (using posterior-adapted class-based weighted fusion) for PAMAP2 and MHEALTH, respectively.

TABLE 3
F1-SCORES FOR SINGLE AND ALL POSSIBLE COMBINATIONS OF ACCELEROMETER SENSORS IN PAMAP2 DATASET

				Posterior-adapted class-based weighted fusion			
	Ankle	**Chest**	**Wrist**	**A+C**	**A+W**	**C+W**	**A+C+W**
Lying	96.88	98.36	90.81	97.25	95.31	95.41	97.06
Sitting	61.65	70.02	82.56	70.00	85.64	85.08	84.24
Standing	72.31	70.55	86.31	72.83	88.98	87.68	87.32
Walking	97.12	82.80	85.70	96.68	96.41	88.65	96.71
Running	89.61	93.41	98.14	98.49	99.19	99.19	99.31
Cycling	94.93	86.27	95.15	96.47	97.58	91.41	98.36
Asc. Stairs	85.00	68.64	44.22	88.01	84.44	71.23	85.88
Desc. Stairs	80.25	77.96	64.01	89.26	88.92	80.06	89.67
Mean	84.72	81.00	80.86	88.62	92.06	87.34	92.32

Of the single location models, classifiers trained on ankle data performed best across both datasets. However, classification accuracy for each PA class varied with accelerometer location. In PAMAP2, the ankle was the best location for walking, ascending stairs, and descending stairs, while the wrist location was best for sitting, standing, running and cycling. The chest location was only best for lying down. In MHEALTH, the ankle location was best for lying down, walking, cycling, and climbing stairs, while the wrist location was best for sitting and standing. The chest location was best for running, and jogging.

TABLE 4
F1-SCORES FOR SINGLE AND ALL POSSIBLE COMBINATIONS OF ACCELEROMETER SENSORS IN MHEALTH DATASET

				Posterior-adapted class-based weighted fusion			
	Ankle	**Chest**	**Wrist**	**A+C**	**A+W**	**C+W**	**A+C+W**
Lying	94.74	89.09	90.85	100.0	96.49	100.0	100.0
Sitting	45.00	44.41	79.23	26.12	81.65	84.93	67.67
Standing	61.02	60.78	86.59	56.54	85.26	86.59	75.95
Walking	95.40	85.94	83.59	97.58	98.70	89.11	94.08
Running	88.62	88.70	76.14	88.78	87.35	86.90	87.71
Cycling	99.36	93.98	94.92	97.32	96.56	99.84	100.0
C. Stairs	98.24	85.58	84.92	98.08	98.72	89.45	94.47
Jogging	86.78	88.47	76.56	89.20	88.41	87.60	88.82
Mean	83.64	79.62	84.10	81.70	91.64	90.55	88.59

Fusion of multiple accelerometer locations using the posterior-adapted class-based decision fusion showed notable improvements in performance compared to the single location models. In PAMAP2, classification performances for the fusion of ankle and wrist accelerometers (A+W) and all three accelerometers (A+C+W) were similar and best among all the combinations. Chest with wrist (C+W) and ankle with chest (A+C) accelerometer locations also exhibited superior performance to that observed for any single location model. In MHEALTH, the best fusion performance was obtained for A+W (91.6%) and C+W (90.6%), with A+C+W also provided outstanding classification performance (88.6%). All combinations except A+C exceeded the performance of any single location model.

D. *Subject-Wise Classification Performance*

Performance differences across different subjects were tested for statistical significance using one-way repeated measures ANOVA. To achieve better statistical confidence, F1-scores for each hold out subject in both datasets were pooled. The results are shown in figure 5. Overall, mean F1-scores differed significantly between the combinations of accelerometer locations (Wilks' Lambda = 0.140, $F(6, 12) = 12.269$, $p < 0.001$). Least significant difference (LSD) post hoc tests revealed a significant improvement in performance when fusing the predictions of two or three accelerometers. All accelerometer combinations except the combination of ankle and chest (A+C) significantly outperformed all single sensor locations. A+W and A+C+W provided the highest average F1-scores across different subjects, but there were not any significant statistical differences between A+W, C+W, and A+C+W.

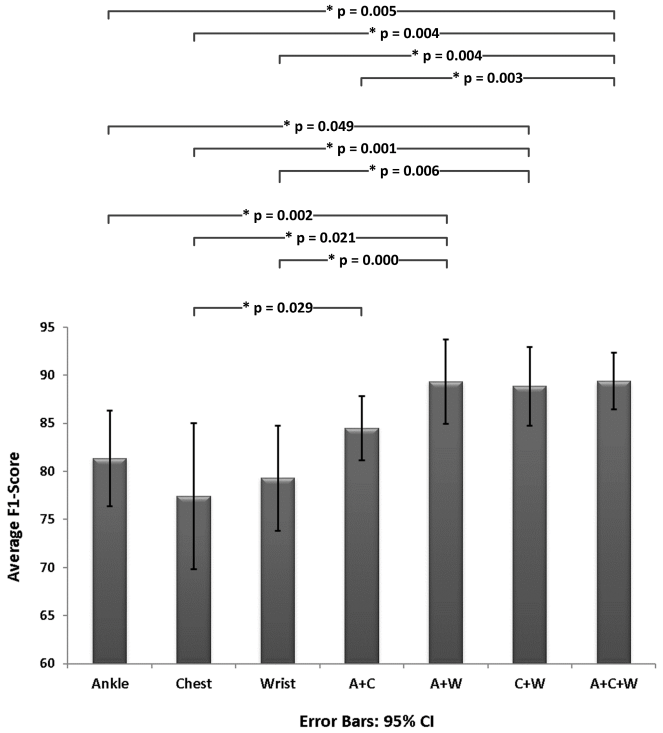

Fig. 5. Average F1-Scores of all single and possible accelerometer combinations across different subjects. Error bars represent 95% confidence intervals. (*) indicates statistical significance ($p < 0.05$)

E. Confusion Matrices

Table 5 and 6 show the confusion matrices of the best-performing accelerometer combination, i.e. combination of ankle and wrist (A+W), using Posterior-adapted class-based weighted fusion for PAMAP and MHEALTH datasets, respectively.

TABLE 5
CONFUSION MATRIX FOR ANKLE AND WRIST COMBINATION (A+W) IN PAMAP2 DATASET

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1 Lying	884	0	1	1	0	0	0	1
2 Sitting	63	701	76	0	0	7	1	2
3 Standing	21	81	763	2	0	3	0	2
4 Walking	0	0	0	1101	0	0	12	5
5 Running	0	0	0	2	430	0	0	3
6 Cycling	0	4	3	1	0	746	1	1
7 Asc. Stairs	0	1	0	49	1	11	342	29
8 Desc. Stairs	0	0	0	10	1	6	21	325

TABLE 6
CONFUSION MATRIX FOR ANKLE AND WRIST COMBINATION (A+W) IN MHEALTH DATASET

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1 Lying	289	0	0	0	0	21	0	0
2 Sitting	0	227	83	0	0	0	0	0
3 Standing	0	18	292	0	0	0	0	0
4 Walking	0	0	0	303	0	0	7	0
5 Running	0	0	0	0	259	0	0	51
6 Cycling	0	1	0	0	0	309	0	0
7 C. Stairs	0	0	0	1	0	0	309	0
8 Jogging	0	0	0	0	24	0	0	286

In both datasets, most misclassifications occurred between similar activity instances, such as misclassification between sitting and standing, and walking and ascending stairs. Most running activities were correctly classified in PAMAP2 dataset, but in MHEALTH, they were misclassified as jogging. Descending stairs was misclassified mostly as ascending stairs.

VI. CONCLUSION

This paper presents a study that investigates the use of multiple accelerometers, placed at three body locations (ankle, chest, and wrist), to effectively identify physical activities. Evaluation was based on two publicly available datasets, namely, PAMAP2 and MHEALTH. The SVM was selected for further analysis as it gave the highest average performance across both datasets. Classification performance depended on both the accelerometer location and activity type. Classifiers trained on ankle data provided the best average performance over all activities. Combinations of classifiers trained on accelerometer data from different locations may improve performance and this was investigated further with model based, class-based and our proposed posterior-adapted class-based weighted decision fusion.

PA recognition using posterior-adapted class-based weighted fusion of multiple accelerometers provided significant improvements in performance in both datasets. Its performance was also found to be better than that observed for model-based, and class-based fusion for all accelerometer combinations. It is consistent with the notion that the combination of ankle and wrist (A+W) accelerometers can capture upper and lower body movements; therefore, can yield significantly higher performance than other combinations. Relative to the two-accelerometer combinations, the addition of the chest location (A+W+C) did not improve PA recognition. Thus, more sensor data does not always result in performance improvements for PA recognition. Considering that chest-mounted accelerometers can be uncomfortable for everyday use; this finding is valuable to motivate future use of ankle and wrist accelerometers for longer-term monitoring of PAs.

A limitation of this paper is that, it uses datasets with only ankle, wrist and chest positioned accelerometers in a controlled setup, and hence overlooks other accelerometer locations such as thigh, hip, etc. For future studies, the proposed framework should be tested by investigating more accelerometer locations, adding more PA classes - especially those that are harder to distinguish, and increase the number of participants to ensure that the findings are generalizable to a wide range of end users. Further test of the proposed method should be done using more PA datasets acquired in different environments to fully study the limitations. This paper has contributed to better understanding of performance improvement with decision fusion in physical activity recognition using multiple accelerometers.

REFERENCES

[1] B. J. Jefferis, P. H. Whincup, L. Lennon, and S. G. Wannamethee, "Longitudinal Associations Between Changes in Physical Activity and

Onset of Type 2 Diabetes in Older British Men The influence of adiposity," *Diabetes Care*, vol. 35, pp. 1876-1883, 2012.

[2] M. Arif and A. Kattan, "Physical Activities Monitoring Using Wearable Acceleration Sensors Attached to the Body," *PloS One*, vol. 10, p. e0130851, 2015.

[3] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, "Activity recognition using a single accelerometer placed at the wrist or ankle," *Medicine and Science in Sports and Exercise*, vol. 45, p. 2193, 2013.

[4] S. J. Preece, J. Y. Goulermas, L. P. Kenney, D. Howard, K. Meijer, and R. Crompton, "Activity identification using body-mounted sensors—a review of classification techniques," *Physiological Measurement*, vol. 30, p. R1, 2009.

[5] R. J. Kate, A. M. Swartz, W. A. Welch, and S. J. Strath, "Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data," *Physiological Measurement*, vol. 37, p. 360, 2016.

[6] M. Hagenbuchner, D. P. Cliff, S. G. Trost, N. Van Tuc, and G. E. Peoples, "Prediction of activity type in preschool children using machine learning techniques," *Journal of Science and Medicine in Sport*, vol. 18, pp. 426-431, 2015.

[7] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition," *Applied Soft Computing*, vol. 37, pp. 1018-1022, 2015.

[8] N. Bicocchi, M. Mamei, and F. Zambonelli, "Detecting activities from body-worn accelerometers via instance-based algorithms," *Pervasive and Mobile Computing*, vol. 6, pp. 482-495, 2010.

[9] K. Ellis, J. Kerr, S. Godbole, J. Staudenmayer, and G. Lanckriet, "Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification," *Medicine and Science in Sports and Exercise*, vol. 48, pp. 933-940, 2016.

[10] A. H. Montoye, J. M. Pivarnik, L. M. Mudd, S. Biswas, and K. A. Pfeiffer, "Comparison of activity type classification accuracy from accelerometers worn on the hip, wrists, and thigh in young, apparently healthy adults," *Measurement in Physical Education and Exercise Science*, vol. 20, pp. 173-183, 2016.

[11] M. J. Mathie, A. C. Coster, N. H. Lovell, and B. G. Celler, "Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement," *Physiological Measurement*, vol. 25, p. R1, 2004.

[12] S. G. Trost, Y. Zheng, and W.-K. Wong, "Machine learning for activity recognition: hip versus wrist data," *Physiological Measurement*, vol. 35, p. 2183, 2014.

[13] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," *IETE Technical Review*, vol. 27, pp. 293-307, 2010.

[14] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, pp. 211-223, 2012.

[15] Y. Sun, M. S. Kamel, and A. K. Wong, "Empirical study on weighted voting multiple classifiers," in *Pattern Recognition and Data Mining*, ed: Springer, 2005, pp. 335-344.

[16] L. Atallah, B. Lo, R. King, and G.-Z. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, pp. 320-329, 2011.

[17] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, ed: Springer, 2004, pp. 1-17.

[18] I. Cleland, B. Kikhia, C. Nugent, A. Boytsov, J. Hallberg, K. Synnes, et al., "Optimal placement of accelerometers for the detection of everyday activities," *Sensors*, vol. 13, pp. 9183-9200, 2013.

[19] N. Kern, B. Schiele, and A. Schmidt, "Multi-sensor activity context detection for wearable computing," in *European Symposium on Ambient Intelligence*, 2003, pp. 220-232.

[20] H. Gjoreski, M. Luštrek, and M. Gams, "Accelerometer placement for posture recognition and fall detection," in *7th International Conference on Intelligent Environments (IE)*, 2011, pp. 47-54.

[21] D. O. Olguın and A. S. Pentland, "Human activity recognition: Accuracy across common locations for wearable sensors," in *10th IEEE International Symposium on Wearable Computers*, Montreux, Switzerland, 2006, pp. 11-14.

[22] F. A. Faria, J. A. Dos Santos, A. Rocha, and R. d. S. Torres, "A framework for selection and fusion of pattern classifiers in multimedia recognition," *Pattern Recognition Letters*, vol. 39, pp. 52-64, 2014.

[23] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, p. 33, 2014.

[24] O. Banos, M. Damas, H. Pomares, F. Rojas, B. Delgado-Marquez, and O. Valenzuela, "Human activity recognition based on a sensor weighting hierarchical classifier," *Soft Computing*, vol. 17, pp. 333-343, 2013.

[25] W. Zhang and Z. Zhang, "Belief function based decision fusion for decentralized target classification in wireless sensor networks," *Sensors*, vol. 15, pp. 20524-20540, 2015.

[26] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, pp. 6474-6499, 2014.

[27] M. Ermes, J. Parkka, J. Mantyjarvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, pp. 20-26, 2008.

[28] S. Pirttikangas, K. Fujinami, and T. Nakajima, "Feature selection and activity recognition from wearable sensors," in *Ubiquitous Computing Systems*, ed: Springer, 2006, pp. 516-527.

[29] S. J. Preece, J. Y. Goulermas, L. P. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 871-879, 2009.

[30] J. Staudenmayer, D. Pober, S. Crouter, D. Bassett, and P. Freedson, "An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer," *Journal of Applied Physiology*, vol. 107, pp. 1300-1307, 2009.

[31] Y. Lin, Q. Hu, J. Liu, J. Chen, and J. Duan, "Multi-label feature selection based on neighborhood mutual information," *Applied Soft Computing*, vol. 38, pp. 244-256, 2016.

[32] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper," in *FLAIRS conference*, 1999, pp. 235-239.

[33] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes," *International Journal of Semantic Computing*, vol. 3, pp. 235-254, 2009.

[34] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, et al., "mHealthDroid: a novel framework for agile development of mobile health applications," in *Ambient Assisted Living and Daily Activities*, ed: Springer, 2014, pp. 91-98.

[35] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, et al., "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *Biomedical Engineering Online*, vol. 14, pp. 1-20, 2015.

[36] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments*, 2012, p. 40.

[37] A. Reiss, M. Weber, and D. Stricker, "Exploring and extending the boundaries of physical activity recognition," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2011, pp. 46-50.